BUILDING BLOCKS

DIFFERENT TYPES OF DATA SOURCES

YORAN VAN OIRSCHOT
Data Engineer

June 28th, 2016

DATA IS EVERYWHERE

These days data is everywhere. You probably heard about exploding data volumes, big data overloads and exponential data growth. Many websites report statistics about data volumes that may blow your mind. However, storing data is useless, unless you can extract value out of it. So where can we find the source of this value?

DIFFERENT TYPES OF DATA SOURCES

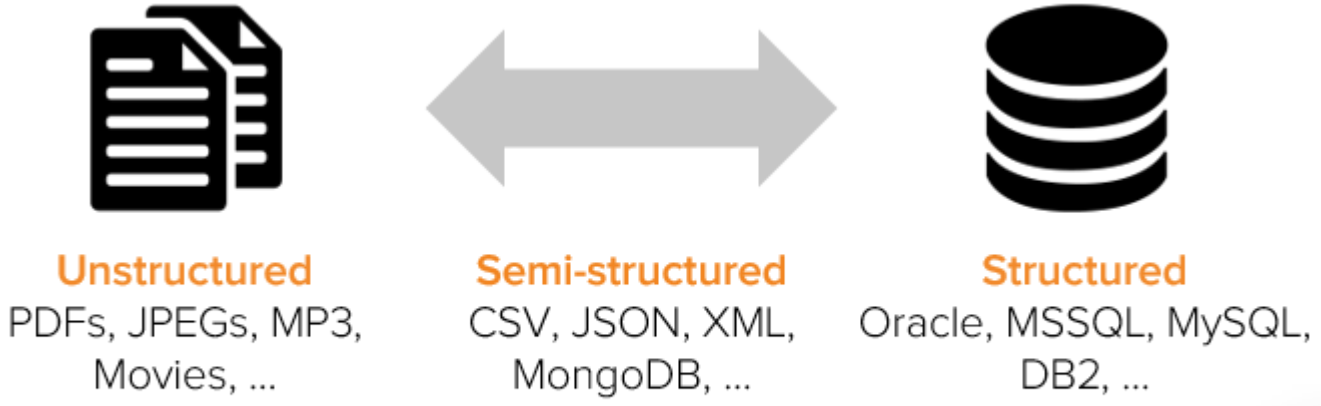
Data manifests itself in many different shapes. Each shape of data may hold much value to the business. In some shapes this is easier to extract than others. Different shapes of data require different storage solutions and should therefore be dealt with in different ways. We at Building Blocks distinguish between three shapes of data:

BUILDING BLOCKS

pictures, sounds or videos. This data is often stored in a repository of files. Think of this as a very well organized directory on your computer's hard drive. Extracting value out of this shape of data is often the hardest. Since you first need to extract structured features from the data that describe or abstract from it. For example, to use text you might want to extract the topics and whether the text is positive or negative about them.

▪ **Structured data** is tabular data (rows and columns) which are very well defined. Meaning that we know which columns there are and what kind of data they contain. Often such data is stored in databases. In databases we can use the power of the language SQL to answer queries about the data and easily create data sets to use in our data science solutions.

▪ **Semi-structured data** is anywhere between unstructured and structured data. A consistent format is defined however the structure is not very strict, like it is not necessarily tabular and parts of the data may be incomplete or differing types. Semi-structured data are often stored as files. However, some kinds of semi-structured data (like JSON or XML) can be stored in document oriented-databases. Such databases allow you to query the semi-structured data.



find them in external data sources like the internet. If you are lucky you will find a data lake that combines all these shapes of data from different sources into a single source. But I will get back on that later.

DATA SOURCES WITHIN THE ORGANIZATION

The first place to look for data is within the organization. Most organizations have an ERP, CRM, Workflow Management, etc. system presently running. These kinds of systems often use a database to store the data in a structured way. These databases contain huge amounts of data from which you can extract value rather easily. For example, from the workflow management system you can easily get insights about bottlenecks in the business processes, or by using data from the ERP system you can make sales predictions.

So far we only looked into structured data sources within the organization. But what about unstructured data? Many organizations receive and send a lot of documents, pictures, sounds or videos. You can probably imagine that, for example, an insurance company receives a lot of claims (on paper or in PDF) possibly attached with pictures. These files are often manually transformed into a more structured format before processing. However, in this transformation some information will be lost. When trying to improve our data science solution we could use these files to extract additional data like the situational sketch. For example, maybe we could improve our fraudulent claim detection using this additional data.

BUILDING BLOCKS

EXTERNAL DATA SOURCES

The real fun starts when we enrich the organizations' data with external data sources. At Building Blocks, we distinguish four kinds of external sources. The most obvious are publicly available datasets. Often governmental organizations release demographic and economic datasets every (few) year(s). An example of such data is the population / km2 per region. We have used such data to improve risk estimation.

There are companies that have made it their core-business to collect, estimate and sell data. We have worked with datasets from such companies and they contain information such as the net income of an address, the size of the house, and even the probability that a person has a dog. We can use this data to enrich an organizations' data to improve their customer profile. Can we use this data to predict the credit risk of our customers?

Many websites these days provide APIs which allow programmers to build interactive apps on their platform (examples are: Twitter, Facebook, LinkedIn, and IMDB). However, such APIs can also be used to collect data. In the case of Twitter, you can request all tweets which contain a certain hash tag. Customer support software are often able to extract social media feeds using these APIs and perform sentiment analysis. Sentiment analysis is a method to determine whether a text is positive or negative about a topic. Using this method, the customer support division can efficiently focus on unsatisfied customers.

Last, but certainly not least is scraping. With scraping you extract relevant data of an unstructured data source. With scraping you are able to extract anything you see on a website.

BUILDING BLOCKS

datasets for sale

Now that I have introduced you to the three shapes of data, a companies' data sources and external data sources, I think we have the basics. In future articles I will go in depth on unstructured data and explain more about the data sources.

RELATED ARTICLES

WHAT IS DATA SCIENCE

Read more

BUILDING BLOCKS

DATA VISUALIZATION: KEEP IT SIMPLE

Read more

BUILDING BLOCKS

Read more

BACK TO OVERVIEW

WANT TO STAY UP TO DATE WITH THE LATEST INNOVATIONS IN DATA SCIENCE?

Subscribe to our newsletter!

Name*

Email address*

Subscribe now!

BUILDING BLOCKS

+31 (0)13 203 2176

AMSTERDAM

Anthony Fokkerweg 1

+31 (0)20 261 9586